Supplementary Online Content

Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, Collins AGE, Frank MJ. Negative symptoms in schizophrenia from a failure in the representation of the expected value of rewards: behavioral and computational modeling evidence. *Arch Gen Psychiatry*. 2012;69(2):129-138.

Appendix. Supplementary materials.

eFigure 1. Patient and control mean (SD) performance (%) on reinforcement learning conditions during the transfer phase immediate testing session.

eFigure 2. Reinforcement learning performance in HNS, LNS, and HC subjects from our previous study.

eTable 1. Transfer Test Phase Performance in Each Group

eTable 2. Measures of Fit for the 3 Different RL Models

This supplementary material has been provided by the authors to give readers additional information about their work.

The following information is included in the Supplementary Materials for this manuscript: 1) Data on patient and control performance in the transfer phase ; 2) A re-analysis of our previous probabilistic selection data (Waltz et al, 2007)[1] analyzing group performance in relation to negative symptom sub-groups, 3) Details on the computational modeling methods and results.

Re-Analysis of Probabilistic Selection Data

We also re-analyzed our previous data on reinforcement learning in schizophrenia using a probabilistic selection task by looking at negative symptom sub-groups to determine whether our previous findings are consistent with those reported in the current manuscript.[1] Negative symptom groups were determined using a median split on the sum of the SANS avolition and anhedonia items. Participants included 24 healthy controls, 16 low negative symptom patients (LNS), and 16 high negative symptom patients (HNS). One-way ANOVA indicated that the 3 groups significantly differed on Choose A performance (the most frequently rewarded stimulus), $F(2, 53) = 7.67$, $p < 0.001$; however, there were no differences among groups on Avoid B performance (the stimulus that was the least rewarding) , $F(2, 53) = 0.37$, $p = 0.69$. Post hoc LSD contrasts indicated that HNS patients chose A significantly less than LNS ($p = 0.006$) or HC ($p < 0.001$) subjects; however, there were no differences between HC and LNS ($p = 0.46$). These findings are consistent with a deficit in Go learning, but intact No Go learning , which is specific to HNS patients. Thus, the re-analysis of our previous data is consistent with our major findings of the current study.

Computational Modeling

The goal of the modeling was to provide a quantitative fit to the pattern of data observed in patients and healthy controls. As described below, we investigated both a standard Actor-Critic architecture and a Q-learning architecture. Neither taken alone could account qualitatively for both healthy control and patient data. We thus investigated a mixture model of Actor-Critic and Q-learning, which leads to better qualitative and quantitative fits for all groups and explains key features of the data, as motivated in the main paper.

Actor-Critic (Basal Ganglia)

According to this model, participants update the expected value V(t) of a state context on each trial t. Each pair of stimuli presented together was represented as a state that might be predictive of the presence of gains or losses. Values are updated as a function of prediction errors using the delta rule:

$$V(s,t+1) = V(s,t) + á_C * ä\ (t),$$

where $á_C$ is the critic learning rate defining the degree to which values are updated on a trial-by-trial basis, and $ä\ (t) = outcome(t)-V(s,t)$ is the reward prediction error showing the discrepancy between expected value V for the current state s and the actual experienced outcome.

Prediction errors in the critic are also used to adjust weights in the actor as follows:

$$w(s,a,t+1)= w(s,a,t) + á_A\ * ä(t),$$

where $w(s,a,t)$ is the stimulus-response weight for the action selected in trial t producing the prediction error $ä\ (t)$ and $á_A$ is the learning rate for the actor which reflects how rapidly its weights are updated. Both learning rates lie in $[0,1]$.

In order to prevent unbound growth of the actor weights, we normalize them by the sum of absolute values, so that they remain on a [-1,1] scale (this also allows proper mixing with Q values, which are naturally bounded, in the hybrid model described below). For example, actor weight for action 1 is normalized according to $w(s,a\_1,t)\ ?\ w(s,a\_1,t) / (|w(s,a\_1,t)| + |w(s,a\_2,t)|)$. To avoid division by a null value we initialized the weights at 0.01. This value is small enough not to bias future probabilities for choosing a stimulus.

Actions are selected according to the standard softmax logistic function:

$$P(a_1,t)= e^{(w(s,a1,t)/â)} / (e^{(w(s,a1\ ,t)/â)} + e^{(w(s,a2,t)\ W/â)}),$$

where a_1 and a_2 denote actions leading to the selection of stimulus 1 or 2 and P(a_1,t) is the probability of choosing action 1. The parameter ? is the softmax temperature and controls the stochasticity of the choice function (e.g. the degree of exploration).

In agreement with previous studies, we also allow positive and negative rewards to be weighed differently. Positive feedback at trial ***t*** was encoded as *outcome(t) = 1-d*, neutral feedback as *outcome(t) = 0* and negative feedback as *outcome(t) = -d*. Thus the free parameter d indicates full neglect of negative outcomes if *d = 0*, full neglect of positive outcomes if *d = 1*, and equal weighing of positive and negative outcomes if *d = 0.5*.

We expected this model to capture both reward learning and loss avoidance, as in prior actor-critic models of avoidance. This model chose randomly at the initial trials of learning, and adjusted the weights associated with a stimulus in the actor following feedback. Weights were increased to reflect learning from positive PEs, and decreased to reflect choices that led to worse-than-expected outcomes. Thus, the model was more likely to repeat choices that led to positive PEs (winners and loss avoiders), while learning to avoid stimuli which produced negative PEs (losers and infrequent winners). It did not, however, distinguish between choices on the basis of actual expected outcome values (gain or loss avoidance, loss or absence of reward).

Q-Learning (OFC)

The Q-Learning model learns the expected value of each action directly, as a function of the prediction error difference between the current expected value of that action and the actual outcome:

$$Q(a,t+1) = Q(a,t) + á_O*(outcome(t) – Q(a,t)),$$

where $á_O$ is the learning rate for the OFC. The Q-Value is only updated for the action selected in the current trial.

Action selection occurs according to the same softmax rule as described higher:

$$P(a_1,t)= e^{(Q(a1,t)/â)} / (e^{(Q(a1,t)/â)} + e^{(Q(a2,t) W/â)}),$$

and the same weighing of positive and negative outcomes through free parameter *d* as for actor-critic is allowed.

As shown in numerous other studies, we expected this model to capture both reward learning and loss avoidance. The model learns the expected values associated with different actions in different states and is thus able to distinguish between choices on the basis of actual outcome values.

Hybrid Actor-Critic Q-Learning Model (OFC-BG interactions)

To account for effects predicted separately by the two previously described models, we propose a hybrid BG-OFC model, in which the BG functions as an actor-critic but its actor values are influenced by top-down OFC Q values. The model includes potentially symmetrical contributions of learned values from both models in the softmax function, by replacing individual contributions of each model by the mixture value:

$$H(s,a1,t)=[(1-c)*w(s,a,t)+c* Q(a,t)]: P(a_1,t)= e^{(H(s,a1,t)/\hat{a})} / (e^{(H(s,a1,t)/\hat{a})} + e^{(H(s,a2,t)/\hat{a})}),$$

where $0 = c = 1$ is a mixing parameter that determines the degree of pure BG vs. OFC contributions. In particular, with c=*0*, the model is reduced to the actor-critic, while with *c=1,* it is reduced to Q-learning. Since both Q-values and normalized weights lie in [-1,1], *c=0.5* indicates equivalent contributions of both systems.

Different model predictions

While all three models predict general reward learning and loss avoidance effects, they each contribute to specific effects observed in the data. In particular,

- The actor-critic model cannot account for sensitivity to actual outcome values, since it only uses reward prediction errors to modify the probability of selecting an action, as opposed to learning specific state action values. On contrary, the Q-learning model predicts sensitivity to actual outcome values, and therefore predicts that subjects will choose a frequent winner over a frequent loss avoider, as seen in HC and LNS subjects.
- The Q-learning model cannot account for the observed preference of frequent loss avoiders (FLA) compared to infrequent winners (IW) across all groups, since infrequent winners have higher expected outcome. In contrast, the AC model can account for this pattern , since frequent loss avoiders lead to frequent positive prediction errors, thus stronger positive actor weights for selecting the loss-avoiding symbol, whereas infrequent winners lead to frequent negative prediction errors, thus negative weights.

Thus, the hybrid model should be able to better account for observed results in patients and healthy subjects. In particular, for a given mixing parameter *c*, the model may recapitulate the preference of healthy subjects to choose frequent winners over frequent loss avoiders (as a function of Q value influences) but still capture a preference for frequent loss avoiders over infrequent winners (due to some influence of AC). Lower *c* values are expected to be associated with the pattern seen in HNS subjects, which correspond to a purer version of AC.

Model fittings

All three models were fitted to subjects' data using the standard likelihood procedure. Specifically, for each participant, we searched for the free parameters that would maximize the likelihood of their own trial-by-trial sequence of choices in both phases of the task, using multiple random starting points.

We found that the hybrid model afforded better fit than both other models for all three groups, even after correction for number of supplementary parameters (3, 4 and 6 respectively for Q-Learning, Actor-Critic and Hybrid models). In the following table, for each group, we report mean (and standard error) pseudo-r2, as well as Akaike Information Criterion (AIC) for complexity penalization.

Results

The fact that all groups benefit from the addition of actor-critic to the Q-learning algorithm is consistent with the finding that all groups showed a preference for frequent loss-avoider trials than infrequent winners. Indeed, this could only be predicted by the AC part of the model (see above).

We performed ANOVAs on the fitted parameter values of the hybrid model, with subject group as a factor. We only found a main effect of group for the mixing parameter c ($F(2,67)=3.8$, $p=.027$; all other parameters $F<2.15$, $p>0.12$). Post-hoc analyses revealed a significantly lower c value for the HNS-SZ patients compared to HC ($t=2.77$, $p=0.008$), as well as a trend for the comparison of LNS-SZ patients to HC ($t=1.7$, $p=0.09$).

Furthermore, the HC group exhibited estimated *c* values that were greater than 0.5 (HC: $c=0.70 \pm 0.06$, $t=3.1$, $p=0.005$; SZ-LNS: $c=0.62 \pm 0.09$, NS). This indicated a greater role for Q-Learning than Actor-Critic in

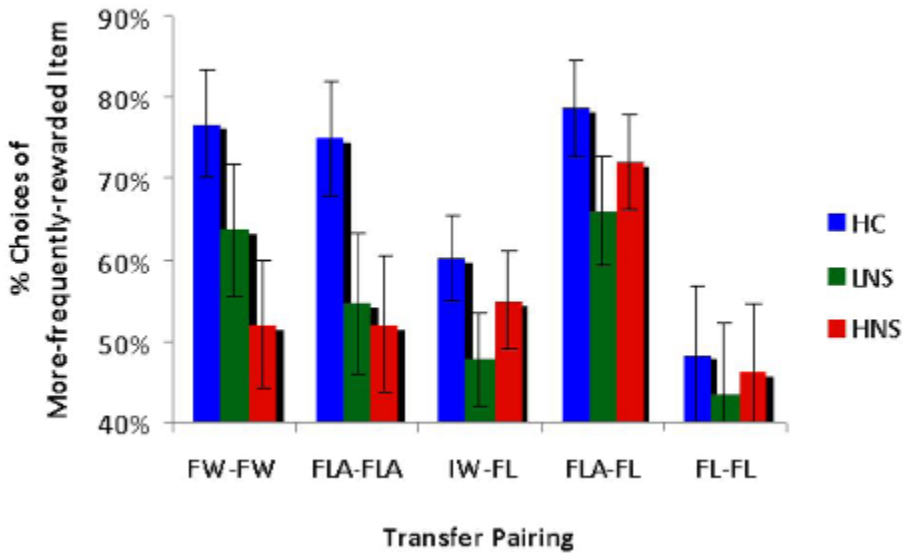
their behavior. Conversely, for the SZ-HNS group the fitted mixing parameter value ($c=0.41 \pm 0.09$) indicated a lesser role for Q-Learning than for Actor-Critic. This is consistent with the observation that those patients do not show a sensitivity to actual outcome value, contrary to HC and SZ-LNS group.

We then used the fitted parameters to simulate the hybrid model for each group, and show that the model can reproduce the key features of the observed data.

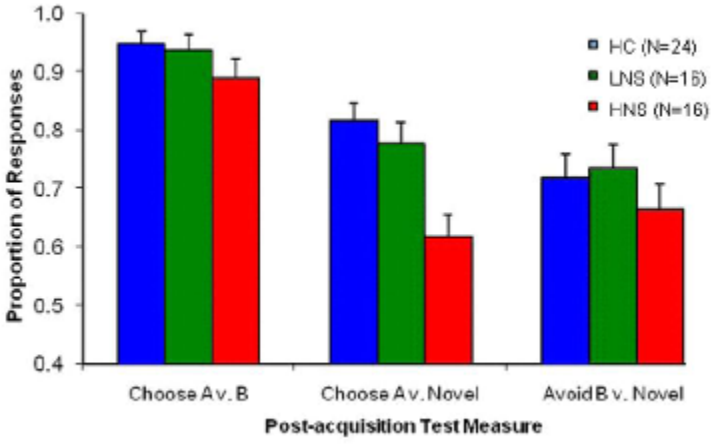
References

1. Waltz JA, Frank MJ, Robinson BM, Gold JM. Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biol Psychiatry.* 2007;62(7):756-764.
2. Moutoussis M, Bentall RP, Williams J, Dayan P. A temporal difference account of avoidance learning. *Network.* 2008;19(2):137-160.
3. Maia TV. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn Behav.* 2010;38(1):50-67.
4. Frank MJ, Claus ED. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol Rev.* 2006;113:300-326.
5. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc Natl Acad Sci U S A*. 2007;104(41):16311-6. Epub 2007 Oct 3.
6. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press; 1998.

Figure Legends



eFigure 1. Patient and Control Mean (SD) Performance (%) on Reinforcement Learning Conditions During the Transfer Phase Immediate Testing Session. Note: FW = Frequent Winner; FLA = Frequent Loss Avoider; FL = Frequent Loser; IW = Infrequent Winner; AB = 90% Gain; CD = 80% Gain; EF = 90% Loss Avoidance; GH = 80% Loss Avoidance.



eFigure 2. Reinforcement Learning Performance in HNS, LNS, and HC subjects from our Previous Study.

eTable 1. Transfer Test Phase Performance in Each Group

	HC (n = 28)	**LNS (n = 22)**	**HNS (n = 25)**	***P* Value**
FW vs. FLA	78 (20)	68 (27)	53 (32)	.005
FW vs. FW	77 (35)	64 (38)	52 (39)	.061
FW vs. FL	89 (16)	86 (18)	79 (25)	.135
FW vs. IW	80 (27)	84 (23)	76 (31)	.594
IW vs. FLA	69 (26)	75 (22)	69 (31)	.660
IW vs. FL	60 (28)	48 (27)	55 (30)	.302
FL vs. FL	48 (46)	43 (42)	46 (43)	.922
FLA vs. FLA	75 (37)	55 (41)	52 (42)	.078
FLA vs. FL	79 (32)	66 (31)	72 (29)	.355
AB Pair	91 (25)	86 (23)	76 (30)	.112
CD Pair	76 (34)	77 (32)	67 (34)	.501
EF Pair	88 (19)	80 (31)	74 (33)	.213
GH Pair	78 (33)	74 (36)	76 (29)	.919

For all pairs, other than the IW vs FLA pair, the values in the table represent the percentage of trials where the participants chose the item with the highest expected value. For the IW vs FLA pair, the value shown is the percentage of choices of the FLA stimulus.

eTable 2. Measures of Fit for the 3 Different RL Models

group	Measure	Hybrid	AC	Q-Learning
HC	pseudo-r2	0.365 (0.039)	0.336 (0.037)	0.311 (0.034)\
	AIC	5528	5774	5996
SZ-LNS	pseudo-r2	0.267 (0.036)	0.2536 (0.034)	0.19876 (0.029)
	AIC	5016	5107	5479
SZ-HNS	pseudo-r2	0.260 (0.033)	0.246 (0.033)	0.190 (0.029)
	AIC	5755	5861	6293